

From Promoter Sequence to Expression: A Probabilistic Framework

By Eran Segal, Yoseph Barash, Itamar Simon, Nir Friedman, and Daphne Koller

Presented at RECOMB 2002

Eugene Ke
Bioinformatics Program
May 29, 2002



Key Points

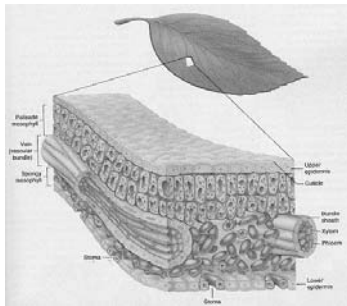
- Biology is in the post-genomic era.
- We can sequence the whole genome or DNA library of organisms.
- The challenge now is to understand how DNA works on a detailed level.
- This paper attempts to model the mechanics of gene expression.
- The model is ambitious as it incorporates data from a multiple of experimental sources.

Central Dogma of Molecular Biology

Central Dogma

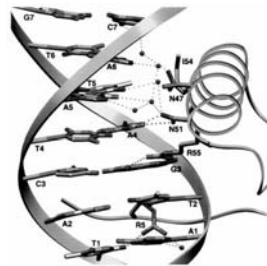


Duplication
Genes are passed on to next generation of cells

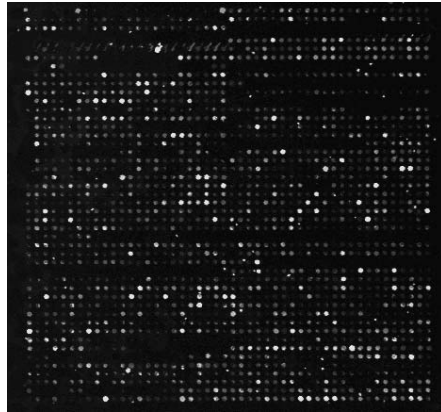


- DNA is a long text composed from a 4-letter alphabet (A,C,G,T).
- Genes are the meaningful portions of DNA.
- DNA is converted into messenger RNA (mRNA) via **transcription**.
- mRNA is used to build proteins, via **translation**.
- Proteins perform all the work in the cell.
- Different cell types perform different functions.
- Therefore, cell types must have different proteins.

Transcription Factors (TFs)

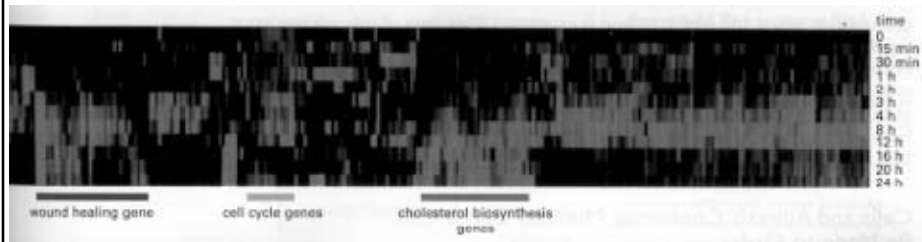


Measuring Gene Expression



- Expression level represents the amount of mRNA present in a cell.
- One DNA array can measure thousands of genes simultaneously.
- Each array is lined with DNA “probes,” for each specific gene.
- mRNA from a cell is extracted from cells and placed on array.
- If a DNA probe responds, corresponding gene is being expressed.

Clustering Expression Levels



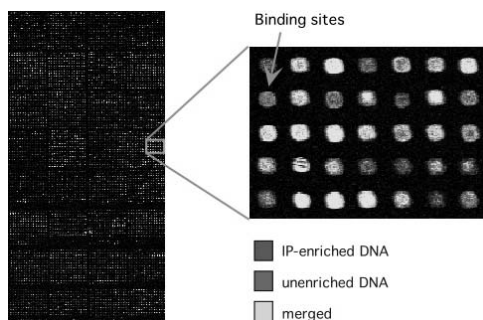
- Using expression data, cluster similarly expressed genes.
 - Genes probably have related function.
- After clustering, we can search promoter regions of clusters.
 - Genes in a cluster are affected by same TFs, therefore will have same promoters.
 - Search promoter regions for similar strings, which is called motifs.
- A motif is a putative promoter.
- Identify probable TFs using motifs.

Finding Promoters Via Sequence

- We now the entire genome of some organisms.
 - We can search directly for motifs.
- Step 1:
 - Search promoter regions of known genes, find motifs.
- Step 2:
 - Group genes by similar motifs.
 - Logic is that if genes are controlled by same TFs, the genes will have similar promoters.
- Step 3:
 - Using databases of known transcription factors, search for probable matches.
- Step 4:
 - Experimentally verify using expression levels of multiple TF combinations.

Complex	Composition	Site name	Site
SBF	Swi6p + Swi4p	SCB	CACGAAA
MBF	Swi6p + Mbp1p	MCB	ACGCGT
Mcm1p	Mcm1p	MCM1	TTACCNAATTTGGTAA
SFF	SFF	SFF	GTMAACAA
Ace2p	Ace2p	SWI5	ACCAGC
Swi5p	Swi5p	SWI5	ACCAGC

Experimentally Finding Binding Sites

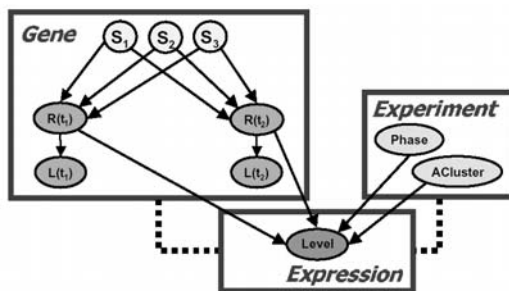


- Localization arrays measure DNA-protein binding.
- They are similar to DNA arrays.
- Run two experiments
- Ratio of intensities show true binding.
- However...
 - Only indicates if TF can bind to promoter, not if TF actually does.
 - Very noisy

Authors' Goals

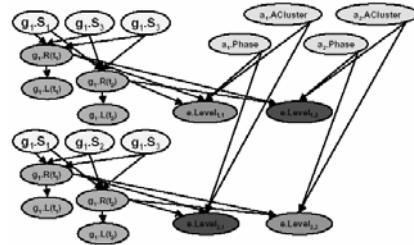
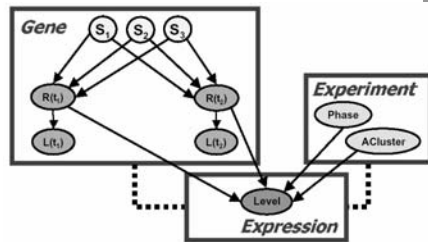
- Analyze two different types of information simultaneously
 - Expression data
 - Sequence data
- Both methods are trying to answer the same question:
 - What genes are co-regulated by the same transcription factors?
- Logically, it is advantageous to combine data.
 - Expression data provides gene expression with respect to time.
 - Sequence data provides hints whether a TF binds to a gene.
- By combining data, it should be possible to determine whether a transcription factor regulates a gene AND under what context.

Probabilistic Relational Model (PRM)



- PRM is an organizational tool.
 - Separate expression and sequence data.
 - Method of relating expression and sequence data.
- Gene
 - Sequence data
 - Localization data
- Experiment & Expression
 - Expression data
- $R(t)$ = Hidden Variable
 - whether a transcription factor regulates a gene

Understanding the gene objects

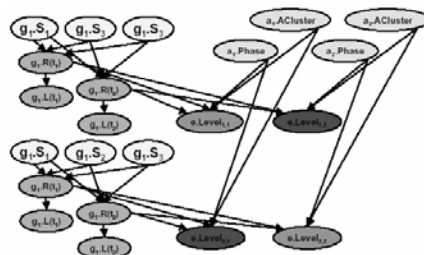
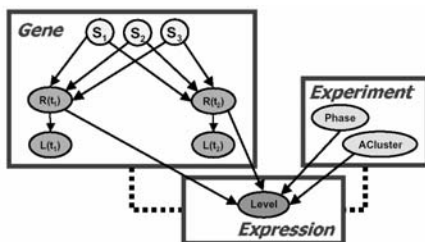


- Transcription Factor t
 - Implicit
 - Enumerated and known at beginning. $t_1 \dots t_m$
 - Described by a Position Specific Scoring Matrix (PSSM)

Gene object g_i

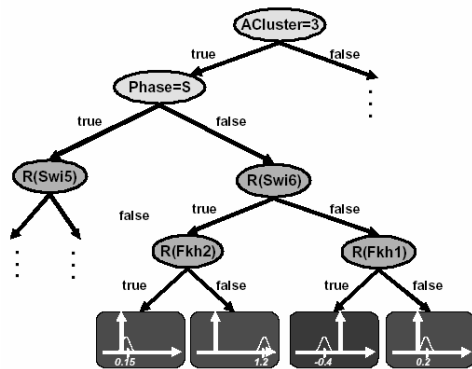
- Contains a promoter region, divided into individual bases $S_1 \dots S_n$
- Contains a Regulates variable $R(t_j)$
 - Whether a TF t_j regulates a gene
 - $R(t_j)$ value for every TF
- May contain Localization variable $L(t_j)$ for a TF t_j

Organizing expression data



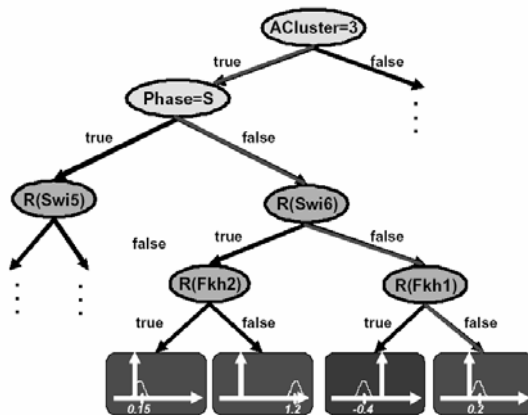
- DNA array a
 - Each array has multiple clusters, called *ACluster*.
 - Each array comes from a specific phase of the cell cycle, denoted by *Phase*
 - Specific to data set
- Expression e
 - Contains expression levels of a gene cluster.
 - *Level* is the expression level of under a specific context.
 - *Array* describes the parent experiment.
 - *Gene* correlates the gene and expression level.

Expression Model



- Expression level depends on three factors
 - Gene cluster
 - Cell-cycle phase
 - TF regulation, $R(t)$
- Dependency is modeled as tree-structured conditional distributions
 - Context specific effects, i.e. phases
 - Combinatorial interactions, such as not $R(\text{Swi6})$ and not $R(\text{Fkh2})$
- Expression levels are shown at leaves
 - Univariate Gaussian distributions

Understanding the Expression Model



- For all genes cluster 3, when they not in the S phase, and are not bound by TF Swi6 nor TF Fkh1 have an expression level centered at 0.2.

Position Specific Scoring Matrix (PSSM)



- Binding sites are “degenerate”
 - Specific but not absolutely so.
- Some mutations in the motif are acceptable while others are not.
- Some position in the motif are highly conserved.
 - In diagram at left, height of letter represents degree of conservation.
- PSSM models acceptable TF binding sites.
 - Each position is represented by a probability of being A,C,G, or T.
- PSSM is a 4xN matrix, where N is the length of the motif.

L(t) is noisy evidence concerning R(t)

- Localization data is labeled as $g.L(t)$.
- Experimental data gives a p-value for each $L(t)$.
- If $R(t)$ is true, we want $L(t)$ to be small.
 - This means we have high confidence as $L(t)$ is a p-value
- If $R(t)$ is false, we want $L(t)$ to be large
 - Data is due to background noise, we have low confidence.
- We assume the probability distribution function is:

$$pdf(L(t) = p \mid R(t) = true) = ce^{-wp}$$

- p is the experimental p-value
- w is an arbitrary weighting factor
- c is a normalizing constant equal to

$$c = \frac{w}{1 - e^{-w}}$$

- In otherwords, $L(t)$ is a noisy sensor, used only as “guidance” for $R(t)$.

Expression Model Learning

- Two main goals of the expression model
 - Learn distributions of expression levels
 - Learn qualitative aspects of the tree structure
- Tree Structure
 - Scoring Function

$$\int \prod_m P(X[m] | V_1[m], \dots, V_n[m], \mathcal{S}_T, \theta_T) P(\theta_T | \mathcal{S}_T) d\theta_T.$$

- Data Set, $D = \{V_1[m], \dots, V_n[m], X[m]\}_{m=1}^M$
- Tree Structure, \mathcal{S}_T
- Gaussian distribution parameters, θ_T
- Greedy local search
 - Trim operation removes nodes
 - Split operation adds nodes

Sequence Model

- In essence, a PSSM is really trying to maximize the probability that a particular substring is nonrandom.
- Sequence Model is given by

$$P(g.R = true | S_1, \dots, S_n) = \text{logit} \left(\log \left(\frac{v}{n-k} \sum_j \exp\{\sum_i w_i[S_{i+j}]\} \right) \right)$$

- With threshold $v = \log \frac{P(g.R=true)}{P(g.R=false)}$
- where $\text{logit}(x) = \frac{1}{1+e^{-x}}$
- $x = \log \left(\frac{P(g.R=true)}{P(g.R=false)} \frac{1}{n-k} \sum_j \prod_{i=1}^k \frac{\psi_i[S_{i+j}]}{\theta_0[S_{i+j}]} \right)$
- Such that the PSSM weights,

$$w_i[l] = \log \frac{\psi_i[l]}{\theta_0[l]}.$$

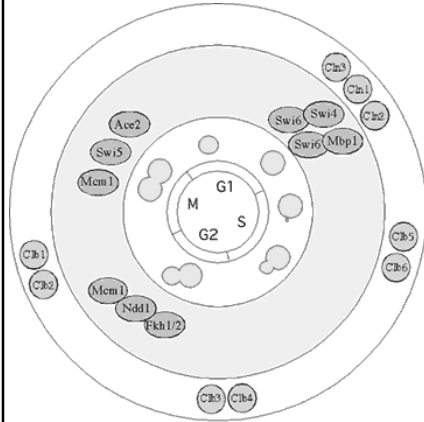
[Learning the Sequence Model]

- Scoring Function
 - Conditional log probability
$$\sum_m \log P(g[m], R(t) | g[m], S_1, \dots, g[m], S_n).$$
 - m is an index denoting a gene.
 - $g[m]$ is a gene in the set of all genes
 - $R(t)$ is whether TF t regulates gene m
 - S is a single nucleotide in the promoter region
- Conjugate Gradient Ascent
 - Trying to find sequences that are over expressed in promoter regions
 - Learn weights of PSSM to optimize scoring function
 - Hill climbing method that finds a local optimum

[Learning the Hidden Variable]

- Regulates variable, $R(t)$, is the hidden variable.
- First, $R(t)$ is initialized using $L(t)$
- Model learns expression data, assigns "hard" values to $R(t)$
- $R(t)$ is further modified using E-M algorithm.
 - As there is a large amount of data, incremental updating.
 - Iterate over all TFs one at a time.
- Main Loop
 - E – step
 - For the current TF t_i , all "hard" values are hidden
 - Assume "hard" values from other TFs are true.
 - Infer a "soft" value using remaining TFs.
 - M – step
 - Use the new "soft" value.
 - Change the parameters in expression and sequence model.
 - Repeat until model converges.

Yeast Cell-Cycle



- The cell-cycle describes the stages a cell undergoes for replication.
- Cell-Cycle has four major phases
 - G1, S, G2, M
- Yeast is a model organism to study cell-cycle control.
 - Many TFs known.
- Large amount of experimental data available.
 - Genome is sequenced
 - Expression data publicly available.
- In figure at left:
 - Inner loop represents the phases
 - Yellow circles represent the physical changes
 - Blue ovals are transcription factors.
 - Outer green circles are cyclins, which are akin to protein “clocks.”

Measuring Predictive Accuracy

- Authors experimented with several model, differing by their combinations of data sources.
- Models trained by cross-validation, and they predict expression level of remaining genes.
- Expression levels of the model are actually Gaussian distributions, with the average and variance as parameters.
- Expression levels of experiments are a range of real values.
- The likelihood function is used to measure how likely the real gene expression levels are given the probability distributions of the model.

$$\text{likelihood}(x_1 \dots x_n; \mu, \sigma^2) = e^{-\sum \frac{(x_i - \mu)^2}{2\sigma^2}}$$

- Average **log** likelihood per gene is used, therefore change in +1 indicates observation twice as likely.
- The closer log likelihood is to zero, the closer the predicted expression level matches experimental expression level.

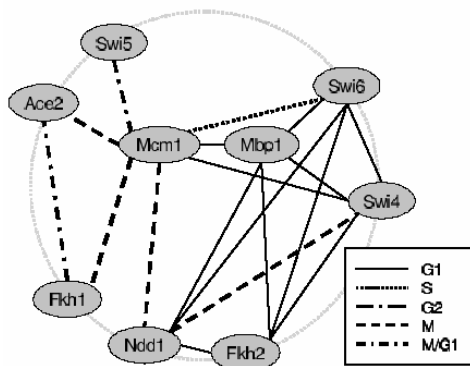
Predictive Results

Model	E	P	R	L	A	S	Average	Variance
1	X	X					-112.24	11.42
2	X	X	X				-134.87	15.29
3	X	X	X	X			-121.48	11.96
4	X	X	X	X	X		-103.76	5.72
5	X	X	X	X	X	X	-94.59	4.13
6	X	X	X		X	X	-95.36	3.90

Results are in log likelihood
 E = expression data
 P = phase data
 R = regulates variable
 L = localization data
 A = acluster
 S = sequence data

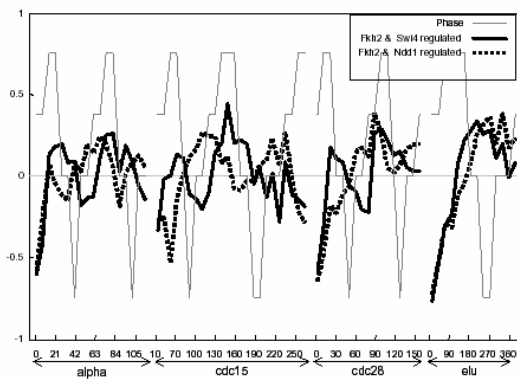
- As more data was included, better performance
- Testing without localization data, gave similar results as with localization data.
- Suggests model can learn promoters from sequence data alone.

Evaluating TF Interactions



- Putative TF interaction maps
 - May interact during specific phases
 - Generated from expression model
- TF interactions previously known
 - Mbp1 & Swi6
 - Swi6 & Swi4
- Interactions missing
 - Ace2 & Swi5
 - Fkh1 & Fkh2
- Novel interactions
 - Ndd1 & Mbp1

Cyclic Gene Expression



- Graph of TF controlled gene expression
 - X-axis is time course
 - Y-axis is expression level
 - Peak of phase graph is G1
 - Trough of phase graphs is G2
- Results are consistent with known biology
 - Swi6 peaks at G1
 - Ndd1 peaks at M

Thoughts...

- Model in theory appears to be a good approach.
- However, difficult to evaluate model due to cryptic results.
 - No comparison to other methods.
- Authors did not experimentally verify their results.
 - It is not sufficient to only search existing data.
 - Need to go to the wet lab.
- Authors are not convincing.
 - Evidence is not conclusive.
 - They admit paper is just a first step.

Summary

- Authors provide very ambitious framework for incorporating large amounts of diverse experimental data.
- Argues that model “learns” really regulation of genes via transcription factors.
- While complex, model appears to be scalable to allow more fewer assumptions.

Suggested References

- General reference
 - *Molecular Biology of the Cell*, by Alberts, et al.
- Publicly available data
 - Expression data
 - <http://genome-www.stanford.edu/cellcycle>
 - Localization data
 - <http://web.wi.mit.edu/young/cellcycle/>
 - Transcription factors
 - <http://transfac.gbf.de/TRANSFAC/>
- Bing Ren, Assistant Professor at UCSD
 - Recently arrived
 - Developed localization array technique
 - <http://ludwig.ucsd.edu/htmls/generegulation.html>